Journal of Machine Learning Research 1–7

# On Atomic Norms

**Arun Tejasvi Chaganty** CHAGANTY@STANFORD.EDU

How many *linear measurements* do you need to (efficiently) recover a low rank matrix? What about a sparse vector or an orthogonal matrix? Given that we know our object of interest has some 'structure', can we answer this question in a general manner?

In this article, I will show you one approach to do so; regression using atomic norms. Most of the material I will cover was presented in the paper, "The Convex Geometry of Linear Inverse Problems" by Venkat Chandrasekaran, et. al.

I will begin by describing the problem of structured recovery with a few motivating examples. I will then review some facts about convex geometry and what atomic norms are in this context. Next, I will show how these properties translate to sharp recovery guarantees, instantiating the framework with two examples; the recovery of low-rank matrices and orthogonal matrices.

There are several more examples described in the paper that I will not talk about. Notably, I will *not* cover the details about theta bodies, approximations that the authors present, which trade off sample complexity for computational efficiency.

## 1. Motivation: Recovering under-determined objects using structure

The problem of recovering sparse but high-dimensional vectors from data ("compressed sensing") has seen recent success in applications to genotyping, medical imaging and comes up commonly when dealing with sensor data. Similarly, the recovery of low-rank matrices and tensors has been used to analyze fMRI data. It also comes up in the recovery of signal from only the magnitude measurements (i.e. without phase), using the PhaseLift algorithm.

### 1.1. The Mixture of Linear Regressions

Personally, I came to be interested in this formulation from my work on applying the method of moments to the mixture of linear regressions, with Percy Liang. In that algorithm, we provided a consistent estimator for the discriminative model where we observe data $(y_i, x_i)$, where $y_i = \beta_h^T x_i + \epsilon_i$ and $\beta_h$ is one of $\beta_1, \beta_2, \ldots \beta_k$; we do not observe which and do not know what the $\beta_h$ are.

The problem is to learn the $\{\beta_h\}_{h=1}^k$, and we do so by observing that $y_i^2$ are linear measurements of $\mathbb{E}_h[\beta_h^{\otimes 2}]$, etc. After recovering these moments, we apply standard techniques to recover $\{\beta_h\}$. In these regression problems, we exploited the low-rank structure to recover the moments with just $O(kd)$ samples instead of potentially $O(d^3)$ samples!

The results of this paper allow us, conceptually at least, to extend this approach to efficiently recover the moments from some linear measurements.

## 2. The Structured Recovery Problem

I will now talk about the exact problem we are looking at. This needs some definitions.

Suppose that the object (matrix, tensor, etc.) we wish to recover is the convex combination of a few "atoms" in the **atomic set** $\mathcal{A}$,

$$x^* = \sum_{i=1}^{k} c_i a_i, \qquad a_i \in \mathcal{A}, c_i > 0.$$

We can define the **atomic norm** of this set to be,

$$\|x\|_{\mathcal{A}} = \inf\{t : x \in t \operatorname{conv} \mathcal{A}\}. \tag{1}$$

Intuitively, this function tells us how much we need to scale $\mathcal{A}$ so that it's convex hull contains $x$. In principle, $\|\cdot\|_{\mathcal{A}}$ is only a norm when $\mathcal{A}$ is "centrally symmetric" ($a \in \mathcal{A} \iff -a \in \mathcal{A}$). The results in the paper extend to non-centrally symmetric $\mathcal{A}$ as well, so we will just assume that $\|\cdot\|_{\mathcal{A}}$ is a valid norm here.

**Examples.** Two standard choices for $\mathcal{A}$ are the set of 1-sparse vectors (for sparse recovery) and set of rank-one matrices (for low rank recovery). The corresponding norms are, as expected, the $L_1$, $\|\cdot\|_1$, norm and the nuclear norm, $\|\cdot\|_*$.

Let $\Phi$ be a linear measurement operator. We wish to ask, how many linear measurements $y = \Phi x^*$ are required to exactly recover $x^*$ from the convex program,

$$\begin{aligned} \hat{x} &= \arg\min_x \|x\|_{\mathcal{A}} \\ \text{subject to } & \; y = \Phi x. \end{aligned} \tag{2}$$

The approach also handles the case where the observations are noisy, $y = \Phi x^* + \omega$, where $||\omega||_2 \leq \delta$;

$$\begin{aligned} \hat{x} &= \arg\min_x \|x\|_{\mathcal{A}} \\ \text{subject to } & \; \|y - \Phi x\|_2 \leq \delta. \end{aligned} \tag{3}$$

In this case, we will ask instead for the number samples required so that $||\hat{x} - x^*|| \leq \epsilon$ for any $\epsilon$.

### 2.1. Convex Geometry Prelimnaries

Before we proceed, I will need to review a few simple but crucial concepts from convex geometry.

The **tangent cone**, $T_{\mathcal{A}}(x)$, gives the directions that *decrease* $\|\cdot\|_{\mathcal{A}}$,

$$T_{\mathcal{A}}(x) = cone\{z - x : \|z\|_{\mathcal{A}} \leq \|x\|_{\mathcal{A}}\}. \tag{4}$$

Similarly, we can define the **normal cone**,

$$N_{\mathcal{A}}(x) = \{s : \langle s, z - x\rangle \leq 0 \ \forall z : \|z\|_{\mathcal{A}} \leq \|x\|_{\mathcal{A}}\}. \tag{5}$$

Finally, the **polar** of a cone $C^*$ is the set of vectors that are obtuse with respect to all vectors in $C$.

$$C^* = \{x \in \Re^p : \langle x, z\rangle \leq 0 \ \forall z \in C\}. \tag{6}$$

Note that the polar cone of the tangent cone is the normal cone, and vice versa.

**Examples.** In the $L_2$ case, i.e. the elements $\mathcal{A}$ have no structural constraints, the tangent cone is the half space, and the normal cone is the normal of this halfspace. In the $L_1$ case, the tangent cone is the cone extending from edges of the $l_1$ ball; the normal cone is the cone bounded by the normals of these two.

## 3. Recovering $x^*$

We have finally come to the meat of this article; in this section, we will study how the properties of the atomic set, $\mathcal{A}$, relate to difficulty of recovering $x^*$.

The first result should illuminate why this approach will work.

**Lemma 1** *(Exact Recovery) $x^*$ is a unique solution to the exact optimization problemEquation (2) iff the tangent cone of the atomic norm at the minimizer and the null-space of the observation operator are transverse,*

$$T_{\mathcal{A}}(x^*) \cup (\Phi) = \{0\}. \tag{7}$$

Recall that the tangent cone $T_{\mathcal{A}}(x^*)$ gives the descent directions of the objective function $(\|\cdot\|_{\mathcal{A}})$ and if $\delta \in (\Phi)$, then $\delta = 0$. Thus, we could move $\delta$ along these directions and reduce the objective *without violating the constraint* $y = \Phi x$.

Now, when we have noisy measurements, we need a stronger condition; that the measurement operator have a "noticeable" projection on the descent directions in $T_{\mathcal{A}}(x^*)$.

**Lemma 2** *(Noisy Recovery) Let $y = \Phi x^* + \omega, \|\omega\| \leq \delta$ be $n$ noisy measurements and $\hat{x}$ be the optimal solution to the noisy optimization problemEquation (3). If, for all $z \in T_{\mathcal{A}}(x^*)$, the measurements are bounded below,*

$$\|\Phi z\| \geq \epsilon \|z\| \qquad\qquad \forall\ z \in T_{\mathcal{A}}(x^*), \tag{8}$$

*then $\|\hat{x} - x^*\| \leq \frac{2\delta}{\epsilon}$.*

**Proof** As $\hat{x}$ minimizes $\|\cdot\|_{\mathcal{A}}$ in the program, we have that $\|\hat{x}\|_{\mathcal{A}} \leq \|x^*\|_{\mathcal{A}}$ and thus that $\hat{x} - x^* \in T_{\mathcal{A}}(x^*)$. Using the linearity of $\Phi$ and the triangle inequality, we derive that,

$$\begin{aligned} \|\Phi(\hat{x} - x^*)\|_{\mathcal{A}} &\leq \|\Phi\hat{x} - y\|_{\mathcal{A}} + \|\Phi x^* - y\|_{\mathcal{A}} \\ &\leq 2\delta, \end{aligned}$$

where the first term is upper bounded by $\delta$ by the optimization program, and the second term is upper bounded by $\delta$ from assumptions on $\omega$.

Finally, using the assumption on $\Phi$, $\|\Phi(\hat{x} - x^*)\|_{\mathcal{A}} \geq \epsilon\|\hat{x} - x^*\|_{\mathcal{A}}$, giving us the finaly result. ■

$\Phi$ is a random quantity, so we will have to show that with sufficient samples, the transversity conditions, (7) and (8), hold with high probability. Note that the assumptions for Lemma 2 subsume those for Lemma 1; it is sufficient for us to estimate a lower bound on

$$\epsilon = \inf_{z \in T_A} (x^*) \frac{||\Phi z||_{\mathcal{A}}}{||z||_{\mathcal{A}}}. \tag{9}$$

This quantity will be referred to as the **minimum gain** of the measurement operator restricted to the cone $T_A(x^*)$.

**Aside: Atomic norms are the "best" convex heuristic.** One intuition to use the atomic norm is as follows. A basic condition we'd like for heuristic penalty is that it be constant for each atom in the set $\mathcal{A}$. Consequently, $a - a'$ should be a descent direction for any $a, a' \in \mathcal{A}$. Requiring that the penalty be convex implies that the set of descent directions $\{a - a' \mid a, a' \in \mathcal{A}\}$ should be a cone. This is precisely the tangent cone at $a \in \mathcal{A}$ with respect to $\text{conv}(\mathcal{A})$.

### 3.1. Gaussian widths and bounding minimum gain

As we noted above, the number of samples required for recovery are determined by minimum gain of $\Phi$. We will now characterize this quantity in terms of the *Gaussian width* of the tangent cone, defined below,

**Definition 3** *The Gaussian width of a set $S \subset \Re^p$ is,*

$$w(S) = \mathbb{E}_{g \sim \mathcal{N}(0, I_n)}[\sup_{z \in S} g^T z].$$

Of course, because of the linearity of $\Phi$, the minimum gain is independent of the length of $z$ so, bounding $||\Phi z||_{\mathcal{A}}$ on $\Omega = T_{\mathcal{A}}(x^*) \cap \mathbb{S}^{p-1}$ is sufficient to bound $||\Phi z||_{\mathcal{A}}$ on $T_{\mathcal{A}}$.

Yehoram Gordon[pdf] proved the following key result on when a random subspace escapes a "mesh" in $\Re^n$, \begin{theorem} Let $\Omega \subset \mathbb{S}^{p-1}$ and $\Phi : \Re^p \to \Re^n$ be a *random map* with i.i.d. standard Gaussian entries. Then,

$$\mathbb{E}[\min_{z \in \Omega} \|\Phi z\|_2] \geq \lambda_n - w(\Omega), \tag{10}$$

where $\lambda_n$ is the expected length of a $n$-dimensional Gaussian vector, $\mathbb{E}_{g \sim \mathcal{N}(0, I_n)}[||g||_2]$. \end{theorem}

The proof follows from a generalization of the Sudakov-Fernique inequality. It is also useful to note that $\lambda_k = \sqrt{2}\Gamma(\frac{k+1}{2})\Gamma(\frac{k}{2})$, which is tightly bounded, $\frac{k}{\sqrt{k+1}} \leq \lambda_k \leq \sqrt{k}$.

### 3.2. Main Theorem: Exponential Convergence to $x^*$

We are finally ready to state and prove the main theorem which shows that we need about $w(\Omega)^2$ samples to exponentially convergence to $x^*$. Theorem 3.1 gives us on the minimum gain that holds in expectation. To extend it to the finite sample regime, we'll exploit the property that $\Phi \to \min_{z \in \Omega} ||\Phi z||_2$ is Lipschitz (with constant 1) and use the concentration of Gaussian measures to show convergence.

**Theorem 4** *Let $\Phi$ be a random map with rows drawn i.i.d. from $\mathcal{N}(0, \frac{1}{n} I_n)$ and $\Omega = T_{\mathcal{A}}(x^*) \cap \mathbb{S}^{p-1}$. Then*

*1. Given measurements $y = \Phi x^*$, then $x^*$ is the unique solution of the the convex program Equation (2) with probability at least $1 - \exp(-\frac{1}{2}(\lambda_n - w(\Omega))^2)$, if $n \geq w(\Omega)^2 + 1$.*

*2. Given measurements $y = \Phi x^* + \omega, \|\omega\|_2 \leq \delta$, then the solution $\hat{x}$ of the the convex program Equation (3) satisfies $\|x^* - \hat{x}\|_2 \leq \frac{2\delta}{\epsilon}$ with probability at least $1 - \exp(-\frac{1}{2}(\lambda_n - w(\Omega) - \sqrt{n}\epsilon)^2)$, if $n \geq \frac{w(\Omega)^2 + 3/2}{(1-\epsilon)^2}$.*

Note that in both cases, the convergence rates are exponential $O(\exp(-n))$.

**Proof** This follows directly by using the property that for any Lipschitz function, with constant $L$,

$$\mathbb{P}[f(g) \geq \mathbb{E}[f] - t] \geq 1 - \exp(-\frac{t^2}{2L^2}). \tag{11}$$

To see that the map $\Phi \to \min_{z \in \Omega} \|\Phi z\|_2$ is Lipschitz, observe,

$$\begin{aligned} \min_{z \in \Omega} \|\Phi z\|_2 - \min_{z' \in \Omega} \|\Phi' z'\|_2 &\leq \|\Phi z'\|_2 - \|\Phi' z'\|_2 && (12) \\ &\leq \|(\Phi - \Phi')\|_F \|z\|_2 && (13) \\ &\leq \|(\Phi - \Phi')\|_F. && (14) \end{aligned}$$

Thus, we have,

$$\mathbb{P}[\min_{z \in \Omega} \|\Phi z\| \geq \epsilon] \geq 1 - \exp(-\frac{1}{2}(\lambda_n - w(\Omega) - \sqrt{n}\epsilon)^2), \tag{15}$$

when $\lambda_n - w(\Omega) - \sqrt{n}\epsilon$. Plugging in $\epsilon = 0$, we get the first condition as well. ■

**Relationship to Radermacher complexity** In the empirical risk minimization framework, the sample complexity is almost entirely defined by the Radermacher complexity, which is analogous to the Gaussian width, but with expectations taken over binary random variables instead of a Guassian.

### 3.3. Properties of Gaussian Widths

The main theorem shows that the sample complexity is pretty sharply bounded by the Gaussian width $w(\Omega)$. The Gaussian width of a set is a very well behaved object with many useful properties. We will use these properties in the applications section, but you should feel free to skip the details here for now.

#### 3.3.1. Relationship to Intrinsic Volume

A large number of properties follow from the fact that $w$ is a valuation, an analogue of a measure. To show this, we'll use a clever fact that an isotropic Gaussian vector can be separated into two independent parts, length and direction. The direction component integrates over the unit sphere.

$$w(S) = \mathbb{E}_g[\sup_{z \in S} g^T z] \tag{16}$$

$$= \underbrace{\lambda_p}_{\text{length}} \int_{\mathbb{S}^{p-1}} \frac{1}{2}(\max_{z \in S} u^T z - \min_{z \in S} u^T z) \underbrace{\mathrm{d}u}_{\text{direction}} \tag{17}$$

$$= \frac{\lambda_p}{2} b(S), \tag{18}$$

where $b(S)$ is the **intrinsic volume** of a set (a deterministic property). It turns out (and this is intuitive) that $\mathrm{d}u$ is a Haar measure, giving us the following properties for free.

- $w(S)$ invariant to translations and unitary transformations.
- Scaling: If $w(tK) \leq tw(K), t > 0$.
- Monotonic: If $S_1 \subset S_2 \subset \Re^p$, $w(S_1) \leq w(S_2)$.
- If $w(S_1 \cup S_2) + w(S_1 \cap S_2) = w(S_1) + w(S_2)$.
- $w(S) = w(\text{conv}(S))$ (because of the sup).
- If $S = S_1 \oplus S_2$, $w(S \cap \mathbb{S}^{p-1}) \leq w(S_1 \cap \mathbb{S}^{p-1}) + w(S_2 \cap \mathbb{S}^{p-1})$.

#### 3.3.2. Other Properties

There are a couple of other cool properties that I won't prove here

- If $V$ is a vector space, $w(V \cap \mathbb{S}^{p-1}) = \sqrt{\dim(V)}$.
- $w(C \cap \mathbb{S}^{p-1}) \leq \mathbb{E}_{g \sim \mathcal{N}(0, I_p)}[\text{dist}(g, C^*)]$ if $C$ is a non-empty convex cone in $\Re^p$. The proof follows from convex duality between $\sup_{z \in C} g^T z$ and $\text{dist}(g, C^*)$.
- A simple corollary; $w(C \cap \mathbb{S}^{p-1}) + w(C^* \cap \mathbb{S}^{p-1}) \leq p$.
- Let $\mathfrak{N}(S, \epsilon)$ be the $\epsilon$-covering number of $S$. Then, by Dudley's inequality,

$$\inf_{\epsilon} c\epsilon \sqrt{\log(\mathfrak{N}(S, \epsilon))} \leq w(S) \leq 24 \int_0^{\infty} \sqrt{\log(\mathfrak{N}(S, \epsilon))} d\epsilon. \tag{19}$$

## 4. Applications

We have covered some fairly complex results in this article. To summarize, we showed the convex optimization problem exponentially converges after getting $O(w(\Omega)^2)$ samples, where $w(\Omega)$ is the Gaussian width of the

tangent cone of the atomic norm $T_{\mathcal{A}}(x^*)$ at the unit sphere. The Gaussian width has several useful compositional properties (just like the Radermacher complexity). To instantiate the framework, we just need to bound the Gaussian width of the atomic set. In this section, we will do so for two examples, orthogonal matrices and low rank matrices.

### 4.1. Orthogonal Matrices

**Lemma 5** *Let $x^*$ be an $m \times m$ orthogonal matrix. Then, letting $\mathcal{A}$ be the set of all orthogonal matrices, $w(T_{\mathcal{A}}(x^*) \cap \mathbb{S}^{m^2-1})^2 \leq \frac{3m(m-1)}{4}$.*

Let's start by defining the atomic norm which should be 1 for all elements in $\mathcal{A}$ and $\leq 1$ for everything in the convex hull $\operatorname{conv} \mathcal{A}$. Note that the set of orthogonal matrices is *not convex*. However, it is not hard to see that every matrix in the convex hull of all $m \times m$ orthogonal matrices has singular values $\leq 1$; let $U$ and $V$ be two orthogonal matrices; $V = AU$, where $A$ is also orthogonal. So,

$$\begin{aligned} \|\theta U + (1-\theta)V\|_2 &= \|(\theta I + (1-\theta)A)U\|_2 \\ &\leq \theta\|I\|_2 + (1-\theta)\|A\|_2 \\ &\leq 1, \end{aligned}$$

where $||U||_2 = \sigma_1(U)$ is the spectral norm, the largest singular value. We have just shown that the atomic norm is the spectral norm!

Using the property that $w$ is rotationally invariant (Property 1 in Section 3.3.1), we just need to consider the tangent cone at $x^* = I$.

$$T_{\mathcal{A}}(I) = \operatorname{cone}\{M - I \mid \|M\|_{\mathcal{A}} \leq 1\}.$$

Every matrix $M$ can be represented as the sum of a symmetric ($A = A^T$) and skew-symmetric matrix ($A = -A^T$), so we can partion $T_{\mathcal{A}}(I)$ into these two spaces. The Lie algebra of skew-symmetric matrices is the tangent space of the orthogonal group $\mathbb{O}(m)$, so we'll have to consider full subspace. Using the symmetry property, this subset is isomorphic to the $\binom{m}{2}$ dimensional vector space that defines the entries above (or below) the diagonal.

The remaining component lies in subspace of symmetric matrices,

$$\begin{aligned} S &= \operatorname{cone}\{M - I : \|M\|_{\mathcal{A}} \leq 1, M \text{ symmetric}\} && (20) \\ &= \operatorname{cone}\{UDU^T - I : \|D\|_{\mathcal{A}} \leq 1, D \text{ diagonal}\} && (21) \\ &= \operatorname{cone}\{U(D-I)U^T : \|D\|_{\mathcal{A}} \leq 1, D \text{ diagonal}\} && (22) \\ &= -\mathrm{PSD}_m, && (23) \end{aligned}$$

where $\mathrm{PSD}_m$ is the set of positive semi-definite $m \times m$ matrices. The positive semidefinte cone is self-dual (i.e. the polar cone $\mathrm{PSD}_m^* = \mathrm{PSD}_m$) and hence contributes (Property 3 in Section (3.3.2)) $\frac{1}{2}\binom{m+1}{2}$.

Putting this together, we have that $w(\Omega)^2 \leq \binom{m}{2} + \frac{1}{2}\binom{m+1}{2}$.

### 4.2. Low-Rank Matrices

**Lemma 6** *Let $x^*$ be a $m_1 \times m_2$ rank-$r$ matrix ($m_1 \leq m_2$). Then, letting $\mathcal{A}$ be the set of unit-norm rank-one matrices, $w(T_{\mathcal{A}}(x^*) \cap \mathbb{S}^{m_1 m_2 - 1})^2 \leq 3r(m_1 + m_2 - r)$.*

Once again, the set of rank-1 matrices is not convex; the sum of two rank-1 matrices is rarely going to stay rank-1! The spectral norm, $||M||_* = \sum_{i=1}^{m} \sigma_i(M)$ (i.e. the sum of singular values) is a suitable atomic norm for this set. Firstly, $\|\cdot\|_*$ is 1 for every element in $\mathcal{A}$ and is also $\leq 1$ for every element in $\operatorname{conv}\mathcal{A}$. While there may be other choices for $\|\cdot\|_{\mathcal{A}}$, the spectral norm is a natural choice to minimize the size of the tangent cone, as we'll shortly see.

Let $x^* = U\Sigma V^T$, where $U \in \Re^{m_1 \times r}$, $\Sigma \in \Re^{r \times r}$ and $V \in \Re^{m_2 \times r}$. The tangent cone at $x^*$ is,

$$T_{\mathcal{A}}(x^*) = \operatorname{cone}\{M - x^* \mid \|M\|_* \leq \operatorname{tr}\Sigma\}.$$

To bound the width of this set, we are going to use Property 2 in Section (3.3.2), which lets us proceed by bounding the distance to the normal cone. The normal cone is described by the sub-differentials at $x^*$,

$$\begin{aligned} N_{\mathcal{A}}(x^*) &= \operatorname{cone}\{UV^T + W : W^T U = 0, WV = 0, \|W\|_{\mathcal{A}}^* \leq 1\} && (24)\\ &= \{tUV^T + W : W^T U = 0, WV = 0, \|W\|_{\mathcal{A}}^* \leq t\}, && (25)\end{aligned}$$

where $||W||_{\mathcal{A}}^*$ is the *dual norm*; in our case it's just the operator norm $\|\cdot\|_2$. Let $\Delta$ be directions in the subspaces of $U$ and $V$ and $\Delta^\perp$ be directions orthogonal to $U$ and $V$.

For any matrix $G$, a point in normal cone is the projection, $||P_{\Delta^\perp}(G)||_2 UV^T + P_{\Delta^\perp}(G)$. Finally,

$$\begin{aligned} w(\Omega) &\leq \mathbb{E}_{G \sim \mathcal{N}((,0),I)}[\operatorname{dist}(G, N_{\mathcal{A}}(x^*))] && (26)\\ &\leq \mathbb{E}[\|G - Z(G)\|_F^2] && (27)\\ &= \mathbb{E}[\|P_\Delta(G) + P_{\Delta^\perp}(G) - \|P_{\Delta^\perp}(G)\|_2 UV^T - P_{\Delta^\perp}(G)\|_F^2] && (28)\\ &\leq \mathbb{E}[\|P_\Delta(G)\|_F^2] + \mathbb{E}[\|P_{\Delta^\perp}(G)\|_2^2]\|UV^T\|_F^2 && (29)\\ &= \mathbb{E}[\|P_\Delta(G)\|^2] + r\,\mathbb{E}[\|P_{\Delta^\perp}(G)\|_2^2]. && (30)\end{aligned}$$

The first term contributes $r(m_1 + m_2 - r)$ because the dimension of $P_\Delta$ is $r(m_1 + m_2 - r)$. The second term is calculated by observing that $P_{\Delta^\perp}(G)$ is an isotropic Gaussian $(m_1 - r) \times (m_2 - r)$ matrix. We have good concentration results on the singular values of such random matrices,

$$\mathbb{P}[\|P_{\Delta^\perp}(G)\|_2 \geq \sqrt{m_1 - r} + \sqrt{m_2 - r} + s] \leq \exp(-s^2/2).$$

It's not hard to show from here that $\mathbb{E}[||P_{\Delta^\perp}(G)||_2^2] \leq (\sqrt{m_1 - r} + \sqrt{m_2 - r})^2 + 2$.

Putting this all together, we get our desired result;

$$\begin{aligned} w(\Omega) &\leq r(m_1 + m_2 - r) + r((\sqrt{m_1 - r} + \sqrt{m_2 - r})^2 + 2)\\ &\leq r(m_1 + m_2 - r) + 2r(m_1 - r + m_2 - 2r + 1)\\ &\leq 3r(m_1 + m_2 - r).\end{aligned}$$

## 5. Conclusions

I hope through this article I've made a case for atomic norms. The framework allows us to reduce the complexity of a recovery problem for arbitrary atomic sets to a Gaussian width computation. Computing the Gaussian widths can still be a hard problem (just like computing the Radermacher complexity), but they provide several useful properties that make these calculations easier. From the few examples I have seen, the Gaussian width calculation requires you to fairly mechanically decompose the tangent cone into spaces that are fairly easy to describe.